GENOME SEQUENCE AS LINEAR UNDIRECTED GRAPH AND ITS DYADIC INTERSECTION WITH THE GRAPHICAL INVERSE

V. Krishnan¹ and A. Reigana Begum²

^{1,2}PG & Research Department of Mathematics, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli -620 020, Tamil Nadu, India Email: ¹vkrishnan1987@gmail.com, ²reigana1992@gmail.com



Abstract

This paper proposes a novel study of genome sequences using graph theoretic tools. A genome sequence is a characteristic sequence of four 'Nucleotides' such as 'Adenine', 'Thymine', Guanine, 'Cytosine'. These four macromolecules are represented by their first letters such as A, T, G and C respectively. Genome is entire set of DNA instructions found in a cell. Human genome consists of 23 pairs of chromosomes located in the cell's nucleus, as well as a small chromosome in the cell's mitochondria. Genome contains all relevant information needed for an individual to grow and function. Human genome contains about 3 billion nucleotides. A sequence of three consecutive nucleotides is called 'Codon', to be more specific, a 'Triplet Codon'. One can construct 64 triplet codons from four nucleotides. During protein synthesis inside a cell, the protein called 'Ribosome' makes use of sequence of triplet codons, chooses amino acids and builds required proteins. These triplet codons play a vital role in the cell action. In a codon sequence, adjacent codons overlap by one nucleotide or by two nucleotides. All hereditary properties of an individual are decided by these overlaps. This paper introduces the structure of overlapping and non-overlapping codons using graph theoretic tools.

1. Introduction

A codon is a DNA or RNA sequence of three nucleotides (a trinucleotide) that forms a unit of genomic information encoding a particular amino acid or signalling the termination of protein synthesis. There are 64 different codons, where 61 specify amino acids and 3 codons such as UAA, UAG and UGA are used as stop signal codons. The question that arises here is as to what extent graph theoretic tools could be used to study DNA/RNA sequences. Traditionally graphs were first used as a purely mathematical way to solve certain problems. One such problem was Königsberg Bridge Problem. 'Path Optimization and Logistics' is another application where graphs are used extensively. Computer networks are probably the easiest thing to represent as a graph since networking uses graphs to represent the layout of any computer network. Applications are many. Almost countless theorems and lemmas have been formulated since 1735. In spite of so much of work carried out in graph theory, it is hardly found in the literature, any application related to the study of DNA/RNA sequences. In fact, DNA/RNA sequences exhibit so much of geometric properties that advocate the effective use of graph theoretical concepts and tools for their analysis. This has been considered as a motivating factor for this research to be carried out.

2. Genetic Code – A Brief Description

Let us denote a finite set of vertices by V. Let us denote the set of edges by E. Then $e_1, e_2, e_3,..., e_n$ are edges, where $E \subseteq \{\{v_i, v_j\} | v_i, v_j \subseteq V \text{ and } v_i \neq v_j\}$. For example, let us consider the 64 triplet codons of a genome sequence. The molecular structures of four macromolecules known as 'Nucleotides' (i) Adenine, (ii) Thymine, (iii) Guanine and (iv) Cytosine are given below. These nucleotides concatenated chemically with what is called as 'Phosphodiester' bond in order to make a genome sequence. A genome sequence is represented as a characteristic sequence consisting of the symbols A (Adenine), T (Thymine), G (Guanine) and C (Cytosine).



Molecular structures of purines and pyrimidines

Phospho-diester bond that builds a Genome Sequence and the double helical Structure of a Genome Sequence are shown in the figure given below. In terms of cellular molecular biology 'genome is a pair of DNA strands of nucleotides entwined in the form of a double helical structure. Each strand is made up of nucleotides chained with the help of 'phosphodiester bonds' and both the strands are entwined with the help of hydrogen bonds. Adenine (A) in one strand weekly pairs up with Thymine (T) in the other strand with the help of hydrogen bond. Similarly, Guanine (G) in one strand weekly pairs up with Cytosine in the other strand with the help of hydrogen bond. Even if the strands of the pair of a genome are separated forcibly inside the cell, they pair up once again with the help of hydrogen bonds. RNA polymerase (a protein inside a cell) makes use of one of the strands of the DNA double helix as template and synthesizes mRNA. This mRNA undergoes several processes and finally becomes a sequence of triplet 'codons'. As per the 'genetic code' dictated by DNA and represented by mRNA, amino acids are added one by one and get translated into polypeptidic sequences.



An RNA (codon sequence) consisting of N nucleotides would contain N/3 triplet codons, N/2 2-tuple codons and N single codons. Thus, one can construct $64^{N/3}$ triplet codon sequences, $64^{N/2}$ 2-tuple codon sequences and 64^{N} single codon sequences. A total of $(64^{N/3}+64^{N/2}+64^{N})$ of RNAs is said to constitute what we call as a finite 'Codon Space'. An infinite 'Codon Space' denoted by the symbol C is potentially realized as the limit of N $\rightarrow \infty$ of $(64^{N/3}+64^{N/2}+64^{N})$. Of all the 64 codons, three are known as termination codons or stop codons, which do not code for any amino acid. They are UAA, UAG and UGA. They act as stop signals for protein synthesis. Remaining 61 codons actually code for 20 amino acids. Refer to figure given below.



The graph $K_{61,20}$ shown above shows 1220 edges. This means there are 1220 correspondences between 61 codons and 20 amino acids. The combinatorics of 1220 correspondences having one at a time, two at a time, and so on up to 1220 at a time amounts to ${}^{1220}C_1 + {}^{1220}C_2 + {}^{1220}C_3 + ... + {}^{1220}C_{1220}$ edges (correspondences). This is called 'Genetic Code'.

20 Amino Acids			
1. alanine - ala - A	6. glutamine - gln - Q	11. leucine - leu - L	16. serine - ser - S
2. arginine - arg - R	7. glutamic acid - glu - E	12. lysine - lys - K	17. threonine - thr - T
3. asparagine - asn - N	8. glycine - gly - G	13. methionine - met - M	18. tryptophan - trp - W
4. aspartic acid - asp - D	9. histidine - his - H	14. phenylalanine - phe - F	19. tyrosine - tyr - Y
5. cysteine - cys - C	10. isoleucine - ile - I	15. proline - pro - P	20. valine - val - V

3. Dyadic Intersection of Linear Undirected Graph and Its Inverse

Here we consider a finite genome sequence as a linear undirected graph with 14 vertices and 13 edges and its graphical inverse as shown below. Let us compare these two graphs, vertices wise, and match the vertices and study the properties.



Let us retain the matching vertices and replace the vertices which are not matching by the symbol #. The resulting graph is shown below. The sequence of matched vertices forms a palindrome sequence.

#	T	 #	#	#	A —	- A	#	#	- #	#	- T	#
						·						

The following proposition holds for any type of alphanumerical sequence.

Proposition 1: "Given any arbitrary numerical or alphanumerical sequence x_1 of length N, one may obtain its graphical inverse as x^{-1} of length N. The sequence x_1 and its inverse x_1^{-1} are pairwise intersected and all matching symbols are identified. Two matched symbols are fused as one. The sequence of all fused matched symbols forms a palindrome sequence".

The above proposition has been verified to be correct by testing 130 virus genome sequences drawn from the NCBI website https://www.ncbi.nlm.nih.gov/. Let us consider the first 70 nucleotides of a virus genome sequence S_1 and its graphical inverse S_1^{-1} . Let us compare the first 70 nucleotides of the virus genome S_1 and its inverse S_1^{-1} .

S ₁	ATT CCG AAAA	GATTGT GTGA	ACCCGTGCGG	TAACCGTCTA	TTTCAAGGAG
	CCAC <mark>TG</mark> GGAC A	GTG <mark>GCC</mark> CGC			
S_1^{-1}	CGCCCGGTGA	CAGGGTCACC	GAGGAACTTT	ATCTGCCAAT	GGCGTGCCCA
	AGTGTGTTAG A	AAAGCCTTA			
S 1 ⁻¹	AGTGTGTTAG A	AAAGCCTTA	GAGGAACIII	ATCIGCCAAI	GGCGIGCCC

The pairwise intersection of S_1 and S_1^{-1} denoted as $S_1 \cap S_1^{-1}$ is **CCGAAGTTGAAGCC**. This sequence could be seen to be a palindrome sequence. The unmatched symbols are denoted by the don't care symbol#. With reference to the figure given below, the resulting pair of sequences is represented as:



where # refers to don't care symbol, OLL refers to 'Open Loop Length', and CLL refers to 'Closed Loop Length'. The value of an OLL is calculated as 2n+1, where 'n' is the number of don't care symbols present in the pair of sequences. The value of a CLL is calculated as 2n+2, where 'n' is the number of don't care symbols present in a single strand. The OLL and CLL values calculated for the above example pair $S_1 \cap S_1^{-1}$ is given as {OLL=7, CLL=8, CLL=4, CLL=6, CLL=78, CLL=6, CLL=4, CLL=8, OLL=7} or simply written as {7, 8, 4, 6, 78, 6, 4, 8, 7}. Note that the value of an OLL is always an odd number and the value of a CLL is always an even number. Note that the sequence pair $S_1 \cap S_1^{-1}$ is an open-ended pair. Now the graphical inverse of the sequence $S_1 \cap S_1^{-1}$ denoted as $(S_1 \cap S_1^{-1})^{-1}$ is CCGAAGTTGAAGCC. It is observed that $S_1 \cap S_1^{-1} \div (S_1 \cap S_1^{-1})^{-1}$, where the symbol \doteq denotes the binary relation 'graphical equivalence', meaning the sequence is a palindrome. Moreover, the sequence of loop values {7, 8, 4, 6, 78, 6, 4, 8, 7} also exhibits palindrome property. In this case, the maximum CLL value is 78 and the minimum CLL value is 4.

Let us consider another sequence S_2 and its inverse S_2^{-1}

S_2	CTTCCG AAAA	GATT <mark>GT</mark> GTGA	ACCCGTGCGG	TAACCGTCTA	TTTCAAGGAG
	CCAC TG GGAC	AGTGGCCCGC			



The pairwise intersection of S_2 and S_2^{-1} denoted as $S_2 \cap S_2^{-1}$ is CCCGAAGTTGAAGCCC. Now the graphical inverse of the sequence $S_2 \cap S_2^{-1}$ denoted as $(S_2 \cap S_2^{-1})^{-1}$ is CCCGAAGTTGAAGCCC. It is observed that $S_2 \cap S_2^{-1} \cdot (S_2 \cap S_2^{-1})^{-1}$, where the symbol \bullet denotes the binary relation 'graphical equivalence'. The sequence CCCGAAGTTGAAGCCC is a palindrome sequence. If the unmatched symbols are replaced by the symbol #, then then the resulting pair of sequences is represented as:



where # denotes don't care symbol and CLL refers to 'Closed Loop Length'. The value of a CLL is calculated as 2n+2, where 'n' is the number of don't care symbols present in the pair of sequences. The CLL values calculated for the above example pair $S_2 \cap S_2^{-1}$ is given as {CLL=6, CLL=8, CLL=4, CLL=6, CLL=78, CLL=6, CLL=4, CLL=6, CLL=6, CLL=4, CLL=6, OLL=6} or simply written as {6, 8, 4, 6, 78, 6, 4, 8, 6}. Note that the sequence pair $S_2 \cap S_2^{-1}$ is a close-ended pair. In this case also, the maximum CLL value is 78 and the minimum CLL value is 4.

4. On the Notion of 'Self Similarity Index (SSI)'

Let us consider the sequence $x_1 = 1,2,3,4,5,6,7,8,9,10$. Now $x_1^{-1} = 10.9.8,7.6,5.4,3.2,1$. Both the sequences are compared. It is observed that not even a single match is found. The number of elements in the sequence is 10. One finds the ratio of '0' to '10' to be '0'. For convenience, one would call this ratio as 'Self Similarity Index (SSI)'.

$$x_1 = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

 $x_1^{-1} = 10, 9, 8, 7, 6, 5, 4, 3, 2, 1$

Proposition 2: One can interpret a discrete Hibert space $l_N: (N \rightarrow \infty)$, which is a sequence space, as a disjoint union of potentially denumerable subspaces, each having a unique SSI value. This proposition is a self-evidential truth.

Proposition 3: Every SSI based subspace of the discrete Hibert space $l_N: (N \rightarrow \infty)$ is potentially denumerable. This proposition is a self-evidential truth.

5. Dyadic Pairing of Nucleotides in a Sequence And Its Inverse

The following proposition holds for any type of genome sequence.

Proposition 4: Given any arbitrary nucleotides sequence x_1 of length N, one may obtain its inverse as x^{-1} of length N. The sequence x_1 and its inverse x_1^{-1} are compared and all of the '**natural pairing**' of nucleotides are identified. The sequence of all such natural pairs of nucleotides $\langle A,T \rangle$, $\langle T,A \rangle$, $\langle G,C \rangle$ and $\langle C,G \rangle$ forms a palindrome sequence.

Let us consider the example sequence S_1 and its inverse S_1^{-1} . Let us carry out natural pairing of the sequence S_1 and its inverse S_1^{-1} and form the sequence of natural pairs as shown below.

S_1	ATTCCGAAAA	GATTGTGTGA	ACCCGTGCGG	TAACCGTCTA	TTT <mark>CA</mark> AGGAG
	CCACTGGGAC	AGTGGCCCGC			
S1-1	CGCCCGGTGA	CAGGGT <mark>CAC</mark> C	GA <mark>GG</mark> AACTTT	ATCTGCCAAT	GGC GT G CC CA
	A <mark>GTG</mark> TGTTAG	AAAAGCCTTA			

The sequence of natural pairs of nucleotides along with don't care symbols is given below.

Now, the sequence of natural pairs of nucleotides along with loop lengths is given below, and this could be seen to be a palindrome sequence.

15<A,T>18<G,C><T,A><G,C>8<C,G><C,G>4<T,A><G,C>8<T,A><A,T>6<C,G><G,C>6<T,A><A,T>8<C, G><A,T>4<G,C><G,C>8<C,G><A,T>C,G>18<T,A>15

Proposition 5: Given any arbitrary nucleotides sequence x_1 of length N, one may obtain its inverse as x^{-1} of length N. The sequence x_1 and its inverse x_1^{-1} are compared and all of the '**unnatural pairing**' of nucleotides are identified. The sequence of all such unnatural pairs of nucleotides <A,C>, <C,A>, <T,G> and <G,T> forms a palindrome sequence.

Let us consider the same example sequence S_1 and its inverse S_1^{-1} . Let us carry out unnatural pairing of the sequence S_1 and its inverse S_1^{-1} and form the sequence of unnatural pairs as shown below.

S_1	ATTCCGAAAA GATTGTGTGA ACCCGTGCGG TAACCGTCTA TTTCAAGGAG CCACTGGGAC AGTGGCCCGC
S1-1	CGCCCGGTGA CAGGGTCACC GAGGAACTTT ATCTGCCAAT GGCGTGCCCA AGTGTGTTAG AAAAGCCTTA

The sequence of unnatural pairs of nucleotides along with don't care symbols is given below.

The sequence of unnatural pairs of nucleotides along with loop lengths is given below, and this could be seen to be a palindrome sequence.

<A,C><T,G>22<T,G><T,G>12<A,C>4<C,A>14<G,T><G,T>6<A,C>10<C,A>6<T,G><T,G>14<A,C>4<C,A>12<G,T><G,T>22<G,T><C,A>

6. Case Study

A Python based computational tool for dyadic intersection of a sequence and its inverse is applied to 130 actual virus genome sequences drawn from NCBI website. All the five propositions stated in this paper were found to be correct. The list of 130 virus genomes is given below.

- S1. >NC 001639.1 Lactate dehydrogenase-elevating virus, complete genome
- S2. >NC 001961.1 Porcine respiratory and reproductive syndrome virus, complete genome
- S3. >NC 002532.2 Equine arteritis virus, complete genome
- S4. >NC_003092.2 Simian hemorrhagic fever virus, complete genome
- S5. >NC_025112.1 Mikumi yellow baboon virus 1 isolate MYBV_M58, complete genome
- S6. >NC_025113.1 Southwest baboon virus 1 isolate SWBV_16986_11/4/2013, complete genome
- S7. >NC 026439.1 Forest pouched giant rat arterivirus isolate PREDICT-06509, complete genome
- S8. >NC 026509.1 DeBrazzas monkey arterivirus isolate PREDICT-06530, complete genome
- S9. >NC_027124.1 Pebjah virus isolate I621, complete genome
- S10. >NC_029053.1 Kafue Kinda chacma baboon virus isolate KKCBV-1, complete genome
- S11. >NC_029992.1 UNVERIFIED: Free State vervet virus isolate VSAI1003, complete genome
- S12. >NC_035127.1 Olivier's shrew virus 1 isolate Gkd-1, complete genome
- S13. >NC_038291.1 Porcine reproductive and respiratory syndrome virus 2, complete genome
- S14. >NC_038293.1 Simian hemorrhagic encephalitis virus isolate Sukhumi, complete genome
- S15. >NC_043487.1 Lelystad virus, complete genome
- S16. >NC_048209.1 Zambian malbrouck virus 1 isolate SHFVagmMal_seqID_01, partial genome

Duplornaviricota

S17. >NC_003555.1 Giardia lamblia virus, complete genome

S18. >NC 005883.1 Chalara elegans RNA Virus 1, complete genome

S19. >NC 007523.1 Coniothyrium minitans RNA virus, complete genome

S20. >NC_009224.1 Botryotinia fuckeliana totivirus 1, complete genome

S21. >NC_009890.1 Black raspberry virus F, complete genome

S22. >NC_014609.1 Armigeres subalbatus virus SaX06-AK20, complete genome

S23. >NC_024151.1 Beauveria bassiana victorivirus NZL/1980 isolate 6887, complete genome

S24. >NC_025214.1 Botryosphaeria dothidea victorivirus 1 isolate GY25, complete genome

S25. >NC_027212.1 Camponotus yamaokai virus genomic RNA, complete genome

S26. >NC_030295.1 Golden shiner totivirus isolate GSTV/US/MN/2014, partial genome

S27. >NC_030867.1 Fusarium poae victorivirus 1 genomic RNA, complete genome

S28. >NC_035674.1 Australian Anopheles totivirus isolate AATV 150840, complete genome

S29. >NC_038928.1 Aspergillus foetidus slow virus 1 CP gene and RdRp gene, genomic RNA

S30. >NC_038929.1 Beauveria bassiana victorivirus 1, complete genome

S31. >NC_040431.1 Diatom colony associated dsRNA virus 13 genomic RNA, complete genome

S32. >NC_040632.1 Gigaspora margarita giardia-like virus 1 isolate GmGlV1-BEG34, complete genome

S33. >NC_040653.1 Fusarium asiaticum victorivirus 1 isolate F16176, complete genome

S34. >NC_040659.1 Diatom colony associated dsRNA virus 10 genomic RNA, complete genome

S35. >NC_040660.1 Diatom colony associated dsRNA virus 11 genomic RNA, complete genome

S36. >NC_040775.1 Diatom colony associated dsRNA virus 12 genomic RNA, complete genome

S37. >NC 040793.1 Alternaria arborescens victorivirus 1 genomic RNA, complete genome

Gresnaviridae

S38. >NC_046959.1 Guangdong greater green snake arterivirus strain LPSG2430 1ab protein, putative glycoprotein, and hypothetical protein genes, complete cds

Kitrinoviricota

S39. >NC_000939.2 Pothos latent virus genes for replicase, capsid protein and movement protein

S40. >NC_001461.1 Bovine viral diarrhea virus 1, complete genome

S41. >NC_001512.1 O'nyong-nyong virus, complete genome

S42. >NC_001513.1 Ononis yellow mosaic virus, complete genome

S43. >NC_001564.2 Cell fusing agent virus strain Galveston, complete genome

S44. >NC_001642.1 Bamboo mosaic virus, complete genome

S45. >NC 001728.1 Odontoglossum ringspot virus, complete genome

S46. >NC_001948.1 Rupestris stem pitting associated virus-1, complete genome

S47. >NC_002604.1 Botrytis virus F, complete genome

S48. >NC_002729.1 Banana mild mosaic virus, complete genome

S49. >NC_002795.1 Aconitum latent virus, complete genome

S50. >NC_003557.1 Garlic latent virus, complete genome

S51. >NC_003603.1 Groundnut rosette virus complete genome, strain MC1

S52. >NC_003608.1 Hibiscus chlorotic ringspot virus, complete genome

- S53. >NC_003634.1 Physalis mottle virus, complete genome
- S54. >NC_003679.1 Border disease virus X818, complete genome
- S55. >NC 003852.1 Obuda pepper virus, complete genome
- S56. >NC_003900.1 Aura virus, complete genome
- S57. >NC_004724.1 Grapevine rootstock stem lesion associated virus, complete genome
- S58. >NC_005062.1 Omsk hemorrhagic fever virus, complete genome
- S59. >NC_005132.1 Botrytis virus X, complete genome
- S60. >NC_006939.1 Olive mild mosaic virus, complete genome
- S61. >NC_007733.2 Angelonia flower break virus, complete genome
- S62. >NC_009028.2 Ilheus virus, complete genome
- S63. >NC_009892.1 Peach chlorotic mottle virus, complete genome
- S64. >NC 011535.1 Grapevine Algerian latent virus, complete genome
- S65. >NC_011538.1 Nemesia ring necrosis virus, complete genome
- S66. >NC_011552.1 Peach mosaic virus, complete genome
- S67. >NC_011559.1 Anagyris vein yellowing virus, complete genome
- S68. >NC_012533.1 Kedougou virus strain DakAar D1470, complete genome
- S69. >NC_012534.1 Bagaza virus strain DakAr B209, complete genome
- S70. >NC_012812.1 Bovine viral diarrhea virus 3 Th/04_KhonKaen, complete genome
- S71. >NC_013006.1 Kalanchoe latent virus, complete genome
- S72. >NC_015782.2 Grapevine Pinot gris virus complete genome, genomic RNA
- S73. >NC 016038.2 Brassica yellows virus isolate BrYV-ABJ, complete genome
- S74. >NC_016404.1 Actinidia virus B, complete genome
- S75. >NC_016440.1 Garlic common latent virus, complete genome
- S76. >NC_016959.1 Ndumu virus, complete genome
- S77. >NC 018713.1 Pestivirus strain Aydin/04-TR, complete genome
- S78. >NC_020470.1 Andean potato latent virus, complete genome
- S79. >NC 020471.1 Andean potato mild mosaic virus isolate Hu, complete genome
- S80. >NC_023439.1 Kama virus strain LEIV-20776Tat polyprotein gene, complete cds
- S81. >NC_023892.1 Gaillardia latent virus isolate 5/18-05-2010, complete genome
- S82. >NC_024458.1 Pitaya virus X isolate P37, complete genome
- S83. >NC 024887.1 Middelburg virus isolate ArB-8422, complete genome
- S84. >NC_026620.1 Jutiapa virus strain JG-128, complete genome
- S85. >NC_028793.2 Phasey bean mild yellows virus isolate NSWCP15, complete genome
- S86. >NC_030693.1 Grapevine Red Globe virus isolate Graciano-T101, complete genome
- S87. >NC_031327.1 Anopheles flavivirus variant 1, complete genome
- S88. >NC 040842.1 Potexvirus sp., complete genome
- S89. >NC 040837.1 Grapevine associated tymo-like virus, complete genome
- S90. >NC_040800.1 Actinidia seed-borne latent virus isolate 01227, complete genome
- S91. >NC 040788.1 Kampung Karu virus isolate SWK P44, complete genome
- S92. >NC 040776.1 Rocio virus strain SPH 34675, complete genome

S93. >NC 039237.1 Bovine viral diarrhea virus 2 polyprotein gene, complete cds

S94. >NC_039218.1 Kyasanur forest disease virus polyprotein gene, complete cds

S95. >NC_039217.1 Phaseolus vulgaris endornavirus 1 isolate PvEV-1_Brazil polyprotein gene, complete cds

S96. >NC_038966.1 Atractylodes mottle virus isolate SK, complete genome

S97. >NC_036587.1 Babaco mosaic virus isolate Tandapi, complete genome

S98. >NC_035462.1 Ocimum basilicum RNA virus 1 isolate DV1 RNA-dependent RNA polymerase and movement protein genes

S99. >NC_035453.1 Actinidia virus 1 isolate K75, complete genome

S100. >NC_035116.1 Lake Sinai Virus TO ORF1, RNA-dependent RNA polymerase, ORF3, and ORF4 genes, complete cds

S101. >NC_035071.1 Apis flavivirus isolate RI-A polyprotein gene, complete cds

S102. >NC_034833.1 Agave tequilana leaf virus isolate agave azul-Mex1, complete genome

S103. >NC_034242.1 Ochlerotatus caspius flavivirus isolate 1608 polyprotein gene, complete cds

S104. >NC 034216.1 Lagenaria siceraria endornavirus-Hubei isolate JZ, complete genome

S105. >NC_034207.1 African eggplant yellowing virus isolate eMA4, complete genome

S106. >NC_034205.1 Grapevine rupestris vein feathering virus isolate Mauzac, complete genome

S107. >NC_033828.1 Peach virus D isolate SK, complete genome

S108. >NC_033725.1 Bamaga virus isolate CY4270 polyprotein gene, complete cds

S109. >NC_033724.1 Kadam virus from Uganda polyprotein gene, complete cds

S110. >NC_033723.1 Gadgets Gully virus from Australia polyprotein gene, complete cds

S111. >NC_033699.1 Jugra virus strain P-9-314 polyprotein gene, complete cds

S112. >NC_033693.1 Bouboui virus strain DAK AR B490 polyprotein gene, complete cds

S113. >NC_032088.1 New Mapoon virus, complete cds

S114. >NC 031752.1 Botrytis cinerea endornavirus 1 strain HBtom -372, complete genome

 $S115. > NC_031463.1\ Cerato basidium\ endormavirus\ B\ isolate\ Murdoch-2\ polyprotein\ and\ ORF2\ genes,\ complete\ cds$

S116. >NC_031462.1 Ceratobasidium endomavirus A isolate Murdoch-1 polyprotein gene, complete cds

S117. >NC_043110.1 Banzi virus strain SAH 336 polyprotein gene, complete cds

Piscanivirinae

S118. >NC 008516.1 White bream virus, complete genome

S119. >NC_026812.1 Chinook salmon bafinivirus isolate NIDO, complete genome

S120. >NC_038295.1 Fathead minnow nidovirus replicase polyprotein 1ab (pp1ab), replicase polyprotein 1a (pp1a), spike glycoprotein (S), membrane protein (M), and nucleocapsid protein (N) genes, complete cds

Remotovirinae

S121. >NC_027199.1 Bovine nidovirus TCH5, complete genome

Serpentovirinae

S122. >NC_024709.1 Ball python nidovirus strain 07-53, complete genome

S123. >NC_033700.1 Xinzhou toro-like virus strain XZSJSC65757 1ab, putative glycoprotein, and hypothetical protein genes,

complete cds

S124. >NC_035465.1 Morelia viridis nidovirus strain S14-1323_MVNV, complete genome

Torovirinae

S125. >NC_007447.1 Breda virus, complete genome

S126. >NC_022787.1 Porcine torovirus strain SH1, complete genome

S127. >NC_034976.1 Goat torovirus strain SZ, complete genome

S128. >NC_046956.1 Bellinger River virus isolate J248, complete genome

S129. >NC_046962.1 Hainan hebius popei torovirus strain LPSC33749 1ab protein, spike protein, and hypothetical protein genes, complete cds

S130. >NC_046963.1 Guangdong red-banded snake torovirus strain LPSF30546 1ab protein, spike protein, hypothetical protein, and putative glycoprotein genes, complete cds

Courtesy: https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi

7. Results and Discussions

Out of 130 virus genome sequences, one hundred and two have been identified as open -ended sequences such as S1, S2, S3, S4, S5, S7, S8, S9, S11, S13, S14, S19, S20, S21, S22, S23, S24, S25, S26, S28, S30, S31, S32, S33, S34, S35, S36, S37, S38, S39, S41, S42, S43, S44, S45, S46, S47, S49, S50, S51, S52, S53, S54, S55, S56, S57, S58, S59, S60, S61, S62, S63, S64, S65, S66, S67, S68, S69, S70, S71, S72, S73, S74, S77, S78, S79, S81, S83, S84, S85, S86, S88, S89, S90, S91, S92, S93, S95, S97, S98, S100, S101, S102, S103, S104, S106, S107, S108, S109, S110, S111, S112, S114, S115, S116, S118, S122, S123, S125, S126, S127, S130 and twenty eight have been identified as close-ended sequences such as S6, S10, S12, S15, S16, S17, S18, S27, S29, S40, S48, S75, S76, S80, S82, S87, S94, S96, S99, S105, S113, S117, S119, S120, S121, S124, S128, S129. The MLL values of 102 open-ended sequences is viewed as a sequence of MLL values and presented as a graph. The maximum and the minimum values in the graph are identified and the difference between the maximum and minimum values is treated as a band. This band is divided into four equal level sub bands. Now all the 102 virus genomes are classified under the four level sub bands. The same procedure is followed to classify 28 close-ended sequences. Table 1 presents values of maximum and minimum loop lengths (MLL) for all open and close -ended sequences and the respective four sub bands.

TD 1	1 1	3 4	•		1	• •	1 (C 1	1	. 1	C		1 1	1	1	1 1	•		
lah	IP I '	1/1	2 V 1m 1	ım an	d m	mmmm	values of		$\mathbf{n} \mid \mathbf{e}$	nathe	ot.	onen	ended	and	CINCE	ended	V11110	(ren(mee
140	юı.	1111	annu	in ai	u III	шшциш	values of		DR	JIEUIS	01	ODCH	unded	anu	CIUSC	Unded	v II U C	2011	JIIICS
										0								0	

<u>S1</u>	Open	Maximum	Minimum	SI	Close	Maximum	Minimum
No	Ended	Values of	Values of	No	Ended	Values of	Values of
110.	Sequences	Loop Lengths	Loop Lengths	110.	Sequences	Loop Lengths	Loop Lengths
1	S1	58	3	1	S6	88	4
2	S2	54	4	2	S10	56	4
3	S3	46	3	3	S12	50	4
4	S4	50	3	4	S15	58	4
5	S5	70	3	5	S16	66	4
6	S7	66	4	-6	S17	52	4
7	S8	52	4	- 7	S18	58	4
8	S9	48	4	8	S2 7	52	4
9	S11	66	4	9	S29	44	4
10	S13	54	3	10	S40	64	4
11	S14	58	3	11	S48	70	4
12	S19	44	4	12	S75	42	4
13	S20	50	4	13	S76	48	4
14	S21	48	4	14	S80	46	4
15	S22	64	4	15	S82	46	4
16	S23	44	4	16	S87	52	4
17	S24	44	4	17	S94	44	4
18	S25	32	4	18	S96	72	4
19	S26	46	4	19	S99	62	4
20	S28	40	4	20	S105	54	4
21	S30	52	4	21	S113	42	4
22	S31	40	4	22	S117	62	4
23	S32	42	4	23	S119	60	4
24	S33	62	4	24	S120	60	4
25	S34	52	4	25	S121	56	4
26	S35	46	4	26	S124	56	4
27	S36	54	4	27	S128	58	4
28	S37	52	4	28	S129	56	4
29	S38	82	4				
30	S39	32	4				

31			
51	S41	48	3
32	\$42	82	1
32	G12	50	
33	\$43	50	4
34	S44	44	4
35	S45	54	4
36	\$46	50	4
50	540	50	7
37	S47	46	4
38	S49	56	4
30	\$50	50	3
37	550	30	5
40	851		4
41	S52	46	4
42	\$53	44	4
42	954	5.4	4
43	534	54	4
44	S55	46	4
45	\$56	80	3
10	057	50	4
40	357	32	4
47	S58	62	4
48	859	44	3
40	\$60	68	4
49	300	08	4
50	S61	40	4
51	S62	56	4
52	\$62	59	1
52	303	50	4
53	S64	48	4
54	S65	54	4
55	\$66	48	4
55	500	0	+
56	S6 7	44	4
57	S68	50	3
58	\$69	48	4
50	507	40	
- 59	S70	40	4
60	S71	50	4
-61	\$72	56	4
()	072	40	
62	5/3	48	4
63	S74	58	3
64	\$77	46	4
65	\$79	56	1
05	3/8	50	4
	870	44	4
66	3/9		
66 67		40	4
66 67 68	\$81 \$83	40	4
66 67 68	S81 S83	40 58	4
66 67 68 69	S81 S83 S84	40 58 50	4 4 4 4
66 67 68 69 70	S81 S83 S84 S85	40 58 50 44	4 4 4 4 4
66 67 68 69 70 71	S81 S83 S84 S85	40 58 50 44 46	4 4 4 4 4 4
66 67 68 69 70 71 72	S81 S83 S84 S85 S86	40 58 50 44 46	$\begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72	S81 S83 S84 S85 S86 S88	40 58 50 44 46 42	4 4 4 4 4 4 4
66 67 68 69 70 71 72 73	S79 S81 S83 S84 S85 S86 S88 S89	40 58 50 44 46 42 46	4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74	S73 S81 S83 S84 S85 S86 S88 S89 S90	40 58 50 44 46 46 42 46 50	4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75	S73 S81 S83 S83 S84 S85 S86 S88 S89 S90 S91 S91	40 58 50 44 46 42 46 50 78	4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91	40 58 50 44 46 42 46 50 78	4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92	40 58 50 44 46 42 46 50 78 50	4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76 77	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93	40 58 50 44 46 42 46 50 78 50 56	4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76 77 78	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95	40 58 50 44 46 42 46 50 78 50 56 52	4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76 77 78 79	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95	40 58 50 44 46 42 46 50 78 50 56 52 36	$ \begin{array}{c} 4 \\ 4 $
66 67 68 69 70 71 72 73 74 75 76 77 78 79	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97	40 58 50 44 46 42 46 50 78 50 56 52 36	$ \begin{array}{c} 4 \\ 4 $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98	40 58 50 44 46 42 46 50 78 50 56 52 36 46	4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S98 S100	40 58 50 44 46 46 50 78 50 56 52 36 46 54	4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100	40 58 50 44 46 42 46 50 78 50 56 52 36 46 50	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101	40 58 50 44 46 42 46 50 78 50 56 52 36 46 52	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102	40 58 50 44 46 42 46 50 78 50 56 52 36 46 52 38	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S103	40 58 50 44 46 46 50 78 50 56 52 36 46 52 38 56	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84	S79 S81 S83 S84 S85 S86 S87 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S103 S104	40 58 50 44 46 42 46 50 78 50 56 52 36 46 52 38 56 38	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 80 81 82 83 84 85	S79 S81 S83 S84 S85 S86 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S104	40 58 50 44 46 42 46 50 78 50 56 52 36 46 52 38 56 38 62	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 62	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S103 S104 S106 S107	40 58 50 44 46 50 78 50 56 52 36 46 51 36 46 52 38 56 38 62 44	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88	S79 S81 S83 S84 S85 S86 S87 S87 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108	40 58 50 44 46 42 46 50 78 50 56 52 36 46 51 36 46 52 38 56 38 62 44 38	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 80 81 82 83 84 85 86 87 88 89	S79 S81 S83 S84 S85 S86 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S107 S108 S109	40 58 50 44 46 42 46 50 78 50 56 52 36 46 51 38 62 44 38 56	$ \begin{array}{c} 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\$
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 92	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 56 56 38 56 46	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110	40 58 50 44 46 50 78 50 56 52 36 46 51 52 38 56 38 56 38 56 44 38 56 48	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91	S79 S81 S83 S84 S85 S86 S87 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S110	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 56 44 38 56 48 56	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92	S73 S81 S83 S84 S85 S86 S87 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S111 S111	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 56 38 56 44 38 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 56 54	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 83	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S111 S111 S111 S111 S111	40 58 50 44 46 46 50 78 50 56 52 36 46 54 52 38 56 38 56 48 56 54	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114	40 58 50 44 46 50 78 50 56 52 36 46 52 38 56 38 56 38 56 54 56 38 56 56 56 56 56 56 56 56 56 56 44 56 54 56 54 56 54 56 54	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 90 91 92 93 94	S79 S81 S83 S84 S85 S86 S87 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S115	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 56 48 56 54 46 48 56 54 46 40	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 90 91 92 93 94 95	S73 S81 S83 S84 S85 S86 S87 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S111 S111 S111 S111 S111 S112 S114 S116	40 58 50 44 46 42 46 50 78 50 56 52 36 46 54 52 38 56 38 56 54 56 56 38 56 54 56 54 56 54 56 54 56 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 86	S73 S81 S83 S84 S85 S86 S87 S88 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S115	40 58 50 44 46 50 78 50 56 52 36 46 54 52 36 46 54 52 38 56 38 56 44 38 56 54 56 54 56 54 56 54 56 54 56 58 58 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96	S79 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S111 S112 S114 S115 S116 S118	40 58 50 44 46 50 78 50 56 52 36 46 52 36 46 51 52 38 56 38 56 38 56 54 56 58 58 58 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 990 91 92 93 94 95 96 97	S73 S81 S83 S84 S85 S86 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S115 S116 S118 S122	40 58 50 44 46 42 46 50 78 50 56 52 36 46 51 38 56 38 56 38 56 38 56 38 56 58 58 58 58 58 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 90 91 92 93 94 95 96 97 98	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S115 S116 S118 S122 S123	40 58 50 44 46 46 50 78 50 56 52 36 46 54 52 36 46 54 52 38 56 38 56 44 38 56 54 56 54 56 54 56 58 58 58 58 58 58 58 58 58 58 58 52 52 53 54 55 56 57 58 52	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 90 91 92 93 94 95 96 97 98 90	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S112 S114 S112 S114 S112 S114 S112 S114 S112 S113 S122 S123	40 58 50 44 46 50 78 50 56 52 36 46 52 36 46 51 52 38 56 38 56 38 56 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $
66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 90 91 92 93 94 95 96 97 98 99	S73 S81 S83 S84 S85 S86 S88 S89 S90 S91 S92 S93 S95 S97 S98 S100 S101 S102 S103 S104 S106 S107 S108 S109 S110 S111 S112 S114 S115 S116 S118 S122 S123 S125	40 58 50 44 46 42 46 50 78 50 56 52 36 46 51 52 38 56 38 56 38 56 48 56 54 56 58 58 58 58 58 58 58 58 58 58 58 58 58 58 58	$ \begin{array}{c} 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ 4\\ $



Fig. 3: Maximum Values of Loop Lengths of Close-Ended Sequences Minimum Values of Loop Lengths of Close-Ended Sequences remains as 4 for all open-ended sequences. 'Maximum Loop Length (MLL)' value has been found to be a reliable quantificational measure of a virus genome, be it open ended or close ended. More the value of MLL of a sequence, less its SSI value. Every virus genome sequence would have a unique SSI value and MLL value. Given a finite set of virus genome sequences, one can find out the MLL value for every virus genome. In the case of 130 virus genome sequences, 102 MLL values have been found for open-ended sequences and 28 MLL values found for closeended sequences. Fig. 4 shows the graph displaying 102 MLL values of open-ended genomes with four levels.



Fig. 4: MLL values of open-ended virus genomes with four levels The genomes S29 and S32 are found to have maximum MLL value 82. Genome S30 is found to have minimum MLL value 32. The band of values from 32 to 82 is further divided into sub bands (i) 32-44, (ii) 45-57, (iii) 58-69 and (iv) 70-82. Fig. 5 shows the four sub bands and virus genomes belonging to those sub bands.

Table 2: Four sub bands and classified open-ended virus genomes Open-Ended Virus

Bands	MLL Ranges	Virus Sequences
1	32-44	\$19, \$23, \$24, \$25, \$28, \$31, \$32, \$39, \$44, \$51, \$53, \$59, \$61, \$67, \$70, \$79, \$81, \$85, \$88, \$97, \$102, \$104, \$107, \$108, \$115
2	45-57	\$2, \$3, \$4, \$8, \$9, \$13, \$20, \$21, \$26, \$30, \$34, \$35, \$36, \$37, \$41, \$43, \$45, \$46, \$47, \$49, \$50, \$52, \$54, \$55, \$57, \$62, \$64, \$65, \$66, \$68, \$69, \$71, \$72, \$73, \$77, \$78, \$84, \$86, \$89, \$90, \$92, \$93, \$95, \$98, \$100, \$101, \$103, \$109, \$110, \$111, \$112, \$114, \$123, \$126, \$127, \$130
3	58-69	S1, S7, S11, S14, S22, S33, S58, S60, S63, S74, S83, S106, S116, S118, S122, S125,
4	70-82	S5 S38 S42 S91

70-82 S5, S38, S42, S91 Fig. 5 shows graph displaying 28 MLL values of close-ended genomes with four levels.



101	S127	54	4	The band o	of values fror	n 32 to 82 is further divided into sub bands (i) 42-		
	S130	54	4	53, (ii) 54- and virus g Table 3: Fo Close-En Bands	-65, (iii) 66-7 genomes be our sub ban nded Virus S MLL Bangass	6 and (iv) 77-88. Fig. 7 shows the four sub bands longing to those sub bands. ds and classified close-ended virus genomes Sequences Virus Sequences		
102				4	1	42-53	\$12, \$17, \$27, \$29, \$75, \$76, \$80, \$82, \$87, \$94, \$113	
							2	54-65
				3	66-76	S16, S48, S96,		
				4	77-88	S6		

- 1. The open-ended virus genomes S19, S23, S24, S25, S28, S31, S32, S39, S44, S51, S53, S59, S61, S67, S70, S79, S81, S85, S88, S97, S102, S104, S107, S108, S115 fall under band 1 and the are likely to undergo structural transformation.
- 2. The close-ended virus sequences S12, S17, S27, S29, S75, S76, S80, S82, S87, S94, S113 fall under band 1 and the are likely to undergo structural transformation.
- 3. The open-ended virus genomes S5, S38, S42, S91 fall under band 4 and they are likely to be stable virus genomes without undergoing any structural transformation.
- 4. The close-ended virus genome S6 falls under band 4 and it is likely to be a stable virus genome without undergoing any structural transformation.

REFERENCES

- 1. "Worlds Record Breaking Plant: Deletes its Noncoding "Junk" DNA". Design & Trend. May 12, 2013. Retrieved 2013-06-04.
- 2. Pennisi, E. (6 September 2012). "ENCODE Project Writes Eulogy for Junk DNA". Science 337 (6099): 1159–1161. doi:10.1126/science.337.6099.1159. PMID 22955811.
- 3. The ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". Nature 489 (7414): 57–74. Bibcode: 2012Natur.489...57T. doi:10.1038/ nature 11247.PMC 3439153. PMID 22955616..
- Costa, Fabrico (2012). "7 Non-coding RNAs, Epigenomics, and Complexity in Human Cells". In Morris, Kevin V. Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection. Caister Academic Press. ISBN 1904455948.
- 5. Carey, Nessa (2015). Junk DNA: A Journey Through the Dark Matter of the Genome. Columbia University Press. ISBN 9780231170840.
- 6. Robin McKie (24 February 2013). "Scientists attacked over claim that 'junk DNA' is vital to life". The Observer.
- Sean Eddy (2012) The C-value paradox, junk DNA, and ENCODE, CurrBiol 22(21):R898–R899.8. Doolittle, W. Ford (2013). "Is junk DNA bunk? A critique of ENCODE". Proc Natl AcadSci USA110 (14): 5294–5300. Bibcode: 2013PNAS..110.5294D. doi:10.1073/pnas.1221376110. PMC 3619371. PMID 23479647.
- Watson, J.D. and Crick, F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, 171, 737-738. https://doi.org/10.1038/171737a0
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA Sequencing with Chain-Terminating Inhibitors. Proceedings of the National Academy of Sciences of the USA, 74, 5463-5467. https://doi.org/10.1073/pnas.74.12.5463
- 10. Southern, E. (1998) Analyzing Polynucleotide Sequences. International Patent Application PCT/GB89/00460.
- Khrapko, K.R., Lysov, Yu.P., Khorlin, A.A., Ivanov, I.B., Yershov, G.M., Vasilenko, S.K., Florentiev, V.L. and Mirzabekov, A.D. (1991) A Method for DNA Sequencing by Hybridization with Oligonucleotide Matrix. DNA Sequence, 1, 375-388. https://doi.org/10.3109/10425179109020793

- 12. Pevzner, P.A. (1989) I-Tuple DNA Sequencing: Computer Analysis. Journal of Biomolecular Structure and Dynamics, 7, 63-73. https://doi.org/10.1080/07391102.1989.10507752
- 13. Bondy, J.A. and Murty, U.S.R. (2008) Graph Theory. Springer, New York.

About the Authors



Dr. V. Krishnan is an Assistant Professor of Mathematics at Jamal Mohamed College, Trichy, with over 15 years of teaching experience. He received his Ph.D. in Mathematics from Srimad Andavan Arts and Science College (2019) and completed post-doctoral research at Pentagram Technologies, Hyderabad (2024). His academic journey includes an M.Sc. (2009), M.Phil. (2010), and B.Sc. in Mathematics (2007), all affiliated with Bharathidasan University. He has published 11 research papers in national and international journals, authored a book chapter (2023), and he actively participates in seminars and workshops. As a life member of IMS, RMS, and MALA, he also supervises four part-time research scholars. His expertise spans topology, probability, operations research, and numerical methods, mathematical model ling and graph theory.



A. Reigana Begum is an Assistant Professor of Mathematics at Jamal Mohamed College, Trichy, with over 8 years of teaching experience. She is pursuing Ph.D. in, Mathematics at Jamal Mohamed College, under guidance of Dr. V. Krishnan. Her academic journey includes an M.Sc. (2015), M.Phil. (2016), and B.Sc. in Mathematics (2013), all affiliated with Bharathidasan University. She has published 7 research papers in national, international journals Her expertise spans Real Analysis, Linear algebra, Operations research and Numerical methods, with research interests in graph theory.

